METHODOLOGY

Power and sample size calculations for Mendelian randomization studies using one genetic instrument

Guy Freeman[1], Benjamin J. Cowling[1], C. Mary Schooling[1,2]

Affiliations:

1. School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 21 Sassoon Road, Pokfulam, Hong Kong Special Administrative Region, China.

2. CUNY School of Public Health at Hunter College, 2180 Third Avenue, New York, NY 10035, USA.

Corresponding author:

Dr Benjamin J Cowling, School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 21 Sassoon Road, Pokfulam, Hong Kong.

Tel: +852 3906 2011; Fax: +852 3520 1945; email: bcowling@hku.hk

Word count (abstract): 177

Word count (main text): 2,719

Running head: Sample size for Mendelian randomization studies

Key words: Mendelian randomization; power; sample size.

FUNDING

This work did not receive any financial support.

ACKNOWLEDGMENTS

We thank Ryan Au Yeung for helpful discussions.

POTENTIAL CONFLICTS OF INTEREST

BJC has received research funding from MedImmune Inc., and consults for Crucell NV.

ABSTRACT

Mendelian randomization, which is instrumental variable analysis using genetic variants as instruments, is an increasingly popular method of making causal inferences from observational studies. In order to design efficient Mendelian randomization studies, it is essential to calculate the sample sizes required. We present formulas for calculating the power of a Mendelian randomization study using one genetic instrument to detect an effect of a given size, and the minimum sample size required to detect effects for given levels of significance and power, using asymptotic statistical theory. We apply the formulas to some example data and compare the results to those from simulation methods. Power and sample size calculations using these formulas should be more straightforward to carry out than simulation approaches. These formulas make explicit that the sample size needed for Mendelian randomization study is inversely proportional to the square of the correlation between the genetic instrument and the exposure and proportional to the residual variance of the outcome after removing the effect of the exposure, as well as inversely proportional to the square of the effect size.

Key messages

- The authors derived formulas to calculate the power and sample size requirements for Mendelian randomization studies using one genetic instrument.

- These formulas permit quicker and easier calculation of sample size compared to simulation approaches.

- The formulas make clear that the sample size requirement for a Mendelian randomization study is inversely proportional to the square of the correlation between the genetic instrument and the exposure and proportional to the residual variance of the outcome after removing the effect of the exposure, as well as inversely proportional to the square of the effect size.

Introduction

It is difficult to make rigorous causal inferences from observational studies due to confounding and reverse causation. One long-standing approach used in the social sciences to tackle such obstacles is the employment of instrumental variables.[1] These only affect the outcome of interest through their effect on the exposure of interest, and are otherwise unconfounded with the outcome. Genetic information has been used in epidemiological studies for more than 20 years[2,3] and is termed Mendelian randomization (MR)[4,5] after Mendel's second law. This law states that the alleles passed on between generations are passed independently of one another (if linkage equilibrium holds true), and more generally of any other factors. The "assignment" of genetic variants for each study subject can therefore be treated as random. Because the instrumental variable is assigned by nature rather than by the researcher, this is a type of natural experiment. While both Mendelian randomization studies and traditional randomized controlled trials are undertaken in order to permit causal inference, they are complementary, providing different information by targeting different causal pathways. An RCT assesses the effect of a modifiable therapy but may not confirm the underlying mechanism, while an MR study assesses the underlying mechanism but does not assess the effect of a therapy.[6]

Before carrying out a trial it is prudent to know how many subjects are required. If too few subjects are recruited then a real effect might not be detected, while recruiting too many will be a waste of resources and potentially expose the participants to unnecessary risk and stress. Sample size calculations are now a standard experimental design procedure in epidemiology,[7] but currently only

simulation approaches have been documented in the literature for MR studies.[8] These approaches are computationally and operationally more expensive to implement than theoretical methods, and subject to simulation error. Here we review how to calculate the effect of an exposure on an outcome – even in the face of confounding – using MR, before proceeding to derive formulas for carrying out power and sample size calculations for MR studies using one genetic instrument.

Estimating causal effects with Mendelian randomization

In order to be able to calculate the causal effect of an exposure on an outcome using MR, certain assumptions must be made.[9] Let X be the random exposure under investigation, Y the outcome variable, and U any unobserved confounders between X and Y. Then for a genetic variant G to be a valid instrumental variable, the following relations must hold true:

1. G and U must not be associated, i.e. the genotype G must not be involved with the confounding between X and Y.
2. G and X must not be independent of each other, i.e. the genotype G must be informative about X.
3. G must only affect Y through its effect on X.

These relations can be represented using a causal graph[9] (Figure 1).

In order to be an instrument G must at a minimum follow the three relations above. However, in order to obtain a point estimate of the causal effect of X on Y, further assumptions must be made about the relations between the variables.

One set of assumptions is that the relations represented in the graph are linear and without interactions, so that (as formulated by Didelez et al[10])

$$E(Y|X = x, U = u) = \beta_{yx}x + \mathrm{h}(u) \quad (1)$$

where h(u) is a function only of u. The aim is to estimate β_{yx}, since this describes the change in Y due purely to X. However, as X and U are correlated, the ordinary least-squares estimator is biased. Instead, a less biased estimator to use here is the Wald estimator[9]

$$\hat{\beta}_{yx} = \frac{\mathrm{cov(Y, G)}}{\mathrm{cov(X, G)}} \quad (2)$$

where cov is the covariance function.

Power and sample size calculations for Mendelian randomization studies with one instrument

In order to undertake power and sample size calculations of MR studies, the distribution of the Wald estimator needs to be known. Assuming the sample is large enough, the distribution of $\hat{\beta}_{yx}$ is approximately[11–13]

$$\hat{\beta}_{yx} \sim N\left(\beta_{yx}, \frac{Var(Y|X)}{n \cdot Var(X) \cdot \rho_{xg}^2}\right) \quad (3)$$

where Var(X) is the variance of the exposure, ρ_{xg} is the correlation between X and G, and Var(Y|X) is the residual variance of Y after removing the effect of X. These can be respectively estimated as the sample variance of the exposure, the

R-squared statistic of a regression of X against G, and the sample variance of the quantities $y - \hat{\beta}_{yx}x$, where $\hat{\beta}_{yx}$ is as given above.

With the estimator $\hat{\beta}_{yx}$ distributed as above, and with a null hypothesis $\beta_{yx} = 0$ and an alternative hypothesis of $\beta_{yx} \neq 0$, the power of detecting a true effect size $\beta_{yx} = \mathrm{b}$ (where b is positive, although by symmetry the power is equivalent for a true effect size of –b) is

$$\text{Power} = 1 + \Phi\left(-z_{\frac{\alpha}{2}} - \frac{\mathrm{b}\rho_{xg}\sqrt{n}}{\sqrt{V}}\right) - \Phi\left(z_{\frac{\alpha}{2}} - \frac{\mathrm{b}\rho_{xg}\sqrt{n}}{\sqrt{V}}\right) \qquad (4)$$

where α is the desired significance level of the test (conventionally 0.05), Φ is the cumulative distribution function of the standard Normal distribution, z_δ is the value which satisfies $\Phi(-z_\delta) = \delta$, and $\mathrm{V} = \frac{Var(Y|X)}{Var(X)}$, the ratio of the sample variance in Y due to factors other than X to the sample variance of X. The derivation of this result is shown in Appendix 1.

If the observed effect size is positive, $\Phi\left(-z_{\frac{\alpha}{2}} - \frac{\mathrm{b}\rho_{xg}^2\sqrt{n}}{\sqrt{V}}\right)$ will be small and can be approximated as zero. By setting the desired power to be 1-β, the formula above can be solved for *n* to give the required sample size,

$$n = \frac{(z_{\frac{\alpha}{2}} + z_{\beta})^2 \cdot V}{b^2 \cdot \rho_{xg}^2} \quad (5)$$

At the usual significance level of $\alpha = 0.05$, $z_{\frac{\alpha}{2}}$ = 1.96. With power set to a conventional 0.8, $z_{\beta} = 0.842$, and the sample size formula becomes

$$n = \frac{7.848 \cdot V}{b^2 \cdot \rho_{xg}^2} \quad (6)$$

For different values of the significance level and the power, the formula will retain the same structure with only the value of 7.848 changing. Note that when $\rho_{xg}^2 = 1$ the minimum sample size required for an MR study is equal to that of an otherwise-identical randomized controlled trial, as the instrument and exposure become perfectly correlated and thus the instrument becomes a perfect proxy for the exposure.

Table 1 shows sample sizes calculated using this formula for a range of other parameter values. Figure 2 shows the parameter combinations that lead to identical sample size requirements. The R code to create Table 1 and Figure 2, as well as the simulation above, is available from the author upon request.

We present an example to help illustrate how to calculate this sample size in practice. As described elsewhere,[14] C-reactive protein (CRP) is a marker for coronary heart disease, as is fibrinogen. The causal pathway between these variables is uncertain. Variations of the *CRP* gene in the form of single nucleotide polymorphisms (SNPs) have been used in Mendelian randomization studies to

learn more about the causality involved,[14,15] as the *CRP* gene is believed to only directly affect CRP. A recent study attempted to assess the causal effect of CRP on fibrinogen using a Bayesian meta-analysis of Mendelian randomization studies.[16] This latter study estimated the causal effect of a unit increase in log(CRP) on fibrinogen to be 0.234 μmol/l using the Wald estimator with the SNP rs1205, although this was not significant at the 95% confidence level. An R^2 between rs1205 and log(CRP) has been observed of around 0.01 [17], while the variance of log(CRP) was found to be 1.11.[14] If these are taken to be the true values of ρ^2_{xg} and V(X), then the only value additionally required to calculate the necessary sample size is the residual variance of fibrinogen after taking into account the causal effect of log(CRP), i.e., V(Y|X). This will depend on the extent to which it is believed that the causal effect of the exposure explains the variation in the outcome. Other sources of variation in the outcome are other covariates and pure random error arising from, for example, measurement error. This concept is analogous to R^2 in the ordinary linear model setting, which gives the proportion of the variation of the outcome explained by the exposure. We assume here that the residual variance of fibrinogen after taking into account the causal effect of log(CRP) is the same as the variance of log(CRP), so that the value of V(Y|X) is also 1.11.

With these values, the required sample size to have a power of 0.8 to detect an effect size of 0.234 at a significance level of 0.05 calculated using the above formula would be around 14,332, because substituting these values in equation (5) leads to the following calculation:

$$\frac{7.848 \times 1}{0.234^2 \times 0.01} \approx 14332 \quad (7)$$

We compare this solution with that given by simulation. We simulated the variables G, U, X and Y, representing the values of the SNP rs1205, confounders, log(CRP) and fibrinogen respectively, according to the distributions given in Appendix 2, which ensured that G would be a valid instrumental variable and that b, ρ_{xg}^2, V(X) and V(Y|X) would be equal to 0.234, 0.01, 1.11 and 1.11

respectively, with varying levels of confounding.

Samples of these variables of size *n* were drawn 25,000 times for varying levels of *n,* with the proportion of the samples where a significant causal effect of X on Y was detected considered to be the power. In order to assess significance, we also simulated samples where there was no causal effect. If the estimate $\hat{\beta}_{yx}$ in the

original sample was smaller than the 2.5% or bigger than the 97.5% percentile of the simulated sampling distribution with no effect then it was considered significant.

With no confounding, a sample size of 14,332 gave a power of around 0.80. With a reasonably low level of confounding between the exposure and outcome – with a correlation between the exposure and confounder of around 0.22 – the sample size required to achieve a power of 0.8 was approximately 14,000, almost exactly the sample size of 14,332 suggested by the formula. At this level of confounding, using a sample size of 14,332 gave a power of around 0.81. As a sensitivity analysis, we also ran a simulation with an extremely high level of confounding

between the exposure and outcome, with a correlation between the exposure and confounder of around 0.97. Sample sizes of around 12,500 were required in this scenario to give power of around 0.8, with some sampling variation around this number. This is around 13% lower than the value of 14,332 given by the formula. Using an *n* of 14,332 in this high confounding simulation gave a power of approximately 0.84.

If the correlation between the exposure protein and the gene were stronger, so that ρ_{xg}^2 were instead equal to 0.03, then the formula would suggest a sample

size of 4778 was required to achieve power of 0.8 at a significance level of 0.05. Simulations with both no confounding and a low level of confounding gave a power of almost exactly 0.8 at this sample size. With a high level of confounding a sample size of around 4200 gave the required power of 0.8, which is around 12% lower than the number given by the formula.

Discussion

In this paper we have presented formulas for calculating the power of a Mendelian randomization study to detect a given effect size and for calculating the required sample size to detect a given effect size with a desired power. Our formulas make explicit that the sample size needed for a Mendelian randomization study is inversely proportional to the square of the correlation between the genetic instrument and the exposure and proportional to the residual variance of the outcome after removing the effect of the exposure, as well as inversely proportional to the square of the effect size.

It must be kept in mind that the formulas provided here should not be applied automatically without forethought. It must be ensured that a Mendelian randomization study is appropriate by carefully considering whether the required assumptions for a valid genetic instrument hold by understanding the biology underlying it. Potential issues with MR studies, and potential remedies, are reviewed in Lawlor et al[4] and Glymour et al.[18]

Even if the basic requirements for a valid MR study are met, the formulas presented above might still not be immediately usable. Firstly, the formulas given here only hold under the assumption that there is a simple linear relationship, without interaction, between the exposure and the outcome. When the outcome is dichotomous, the formulas in this paper can still be used as long as the linearity assumptions they depend on are not deviated from too much. They would therefore only make good approximations if the assumption that the outcome changes linearly with the exposure holds well over the range of the exposure, which is more likely to be the case if both the causal effect and the range of the exposure are relatively small.[10] In general, though, non-linearity complicates calculation of the causal effect so that the Wald estimator is not valid any longer and the formulas given here would no longer be applicable.

Secondly, the formulas rely on the asymptotic distribution of the instrumental variable estimator as given in Equation 3. In any finite sample, though, the actual sampling distribution will differ from this to a certain extent.[19] The extent of this difference determines the extent to which the asymptotic distribution in

Equation 3 is a good approximation to use. It is well known that a weak instrument – i.e., an instrument with a very low correlation with the exposure of interest relative to the confounding present between the exposure and the outcome – can lead to a substantially different estimate of the causal effect from the true value.[12,19–21] In the simulation example above, it was seen that when confounding was extreme the sample size required differed more from the one calculated with the formula compared to a scenario where there is no or low confounding. When the instrument was stronger, the discrepancy between the sample size the formula suggested and that yielded by simulation was slightly smaller. Although the formula will always provide an initial estimate, the formula performs better in situations without extreme confounding.

With the widespread availability of genetic data, it is becoming increasingly common to perform Mendelian randomization using multiple genetic variants. While the formulas here were derived for the case where there is one genetic variant, they can be extended to the case where there are multiple genetic variants. This is achieved through the creation of an "allele score", which is a weighted or unweighted sum of the number of alleles in the genetic variants that are considered to have a positive effect on the exposure of interest,[8,22] which is then used in turn as the genetic instrument. If the individual genetic variants have similar effects on the exposure, then the unweighted allele score has been shown to lead to a robust causal estimate of the exposure on the outcome, with slightly lower power being made up for by lower bias compared to estimating the effect of each genetic variant on the exposure separately when using, for example, the two-stage least squares estimator.[8,23] This allele score can then be

treated in the same way as a single genetic instrument is handled in this paper, i.e. ρ_{xg} becomes the correlation between the exposure and the allele score.

We showed how the sample size required for a Mendelian randomization study can also be found using simulation instead of the formulas derived here. This has the advantage of avoiding the finite-sample bias our formulas suffer from due to depending on asymptotic statistical theory. However, the disadvantages are that it is not trivial to set up the required simulation so that the variables under consideration have the desired distributions, and that while it is relatively simple to find the power for a particular sample size (subject to simulation error), to find the sample size required to achieve a fixed power it is necessary to run a simulation for different sample sizes until the desired power is achieved. We therefore recommend our formulas as a useful and fast approximate calculator for power and sample size requirements.

In conclusion, we have derived and provided a procedure for calculating without simulation the sample size required to carry out a Mendelian randomization study with a desired significance level and power. Using an analytic formula also highlights some points that can facilitate design of the most efficient Mendelian randomization study. First, the Mendelian randomization sample size calculation only differs from that of an RCT in that the sample size is inversely proportional to the square of the correlation between the genetic variants and the exposure, making plain the importance of choosing a genetic instrument that is more strongly correlated with the exposure. Second, the proportion of the expected

variation in the outcome that is explained by the exposure alone affects the sample size, just as it does for an RCT.

REFERENCES

1. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Amer Statist Assoc*. 1996 Jun 1;**91**(434):444–455.

2. Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet*. 1986 Mar;**327**(8479):507–508.

3. Davey Smith G, Ebrahim S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. 2003 Feb 1;**32**(1):1–22.

4. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Stat Med*. 2007 Sep 20;**27**(8):1133–1163.

5. Wehby GL, Ohsfeldt RL, Murray JC. 'Mendelian randomization' equals instrumental variable analysis with genetic instruments. *Stat Med*. 2008;**27**(15):2745–2749.

6. Schooling CM, Freeman G, Cowling BJ. Mendelian Randomization and Estimation of Treatment Efficacy for Chronic Diseases. *Am J Epidemiol* [Internet]. 2013 Apr 12 [cited 2013 Apr 23]; Available from: http://aje.oxfordjournals.org/content/early/2013/04/12/aje.kws344

7. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *N Engl J Med*. 1978 Sep 28;**299**(13):690–694.

8. Pierce BL, Ahsan H, VanderWeele TJ. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Int J Epidemiol*. 2011 Jun 1;**40**(3):740–752.

9. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res*. 2007;**16**(4):309–330.

10. Didelez V, Meng S, Sheehan NA. Assumptions of IV methods for observational epidemiology. *Stat Sci*. 2010;**25**(1):22–40.

11. Wooldridge JM. Introductory econometrics: a modern approach. 4th ed. South-Western College Pub; 2008.

12. Nelson CR, Startz R. The Distribution of the Instrumental Variables Estimator and Its t-Ratio When the Instrument is a Poor One. *J Bus (Chic Ill)*. 1990 Jan 1;**63**(1):S125–S140.

13. Martens EP, Pestman WR, Boer A de, Belitser SV, Klungel OH. Instrumental Variables. *Epidemiology*. 2006 May;**17**(3):260–267.

14. C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC). Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ*. 2011 Feb 15;**342**(feb15 2):d548–d548.

15. Keavney B, Danesh J, Parish S, et al. Fibrinogen and coronary heart disease: test of causality by 'Mendelian randomization'. *Int J Epidemiol*. 2006 Aug 1;**35**(4):935–943.

16. Burgess S, Thompson SG. Bayesian methods for meta-analysis of causal relationships estimated using genetic instrumental variables. *Stat Med*. 2010;**29**(12):1298–1311.

17. Burgess S, Thompson SG, CRP CHD Genetics Collaboration. Methods for meta-analysis of individual participant data from Mendelian randomisation studies with binary outcomes. *Stat Methods Med Res* [Internet]. 2012 Jun 19

[cited 2013 Mar 20]; Available from: http://smm.sagepub.com/content/early/2012/06/18/0962280212451882

18. Glymour MM, Tchetgen Tchetgen EJ, Robins JM. Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *Am J Epidemiol*. 2012 Feb 15;**175**(4):332–339.

19. Nelson CR, Startz R. Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica*. 1990 Jul 1;**58**(4):967–976.

20. Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *J Amer Statist Assoc*. 1995 Jun 1;**90**(430):443–450.

21. Sheehan NA, Didelez V. Commentary: Can 'many weak' instruments ever be 'strong'? *Int J Epidemiol*. 2011 Jun 1;**40**(3):752–754.

22. Palmer TM, Lawlor DA, Harbord RM, et al. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Stat Methods Med Res*. 2012 Jun 1;**21**(3):223–242.

23. Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization [Internet]. 2012. Available from: http://www.phpc.cam.ac.uk/ceu/files/2012/11/scoreone140612-1.pdf

24. DasGupta A. Probability for statistics and machine learning. Springer; 2011.

Appendix 1

The Wald estimate $\hat{\beta}_{yx}$ is normally distributed with mean β_{yx} and variance

$\frac{Var(Y|X)}{n\rho_{xg}^2 Var(X)}$. Under the null hypothesis of X having no causal effect on Y, i.e. $\beta_{yx} = 0$,

the null hypothesis will be rejected at significance level α when

$|\hat{\beta}_{yx}| > z_{\frac{\alpha}{2}} \cdot \frac{\sqrt{V}}{\rho_{xg}\sqrt{n}}$. If the true value of β_{yx} is b, then the probability of rejecting the

null hypothesis, i.e. the power of the study, is

$$\mathrm{P}\left(\left|\hat{\beta}_{yx}\right| > z_{\frac{\alpha}{2}} \cdot \frac{\sqrt{V}}{\rho_{xg}\sqrt{n}}\right) = \mathrm{P}\left(\hat{\beta}_{yx} < -z_{\frac{\alpha}{2}} \cdot \frac{\sqrt{V}}{\rho_{xg}\sqrt{n}}\right) + \mathrm{P}\left(\hat{\beta}_{yx} > z_{\frac{\alpha}{2}} \cdot \frac{\sqrt{V}}{\rho_{xg}\sqrt{n}}\right)$$

$$= \mathrm{P}\left(\frac{\rho_{xg}\sqrt{n}(\hat{\beta}_{yx} - b + b)}{\sqrt{V}} < -z_{\frac{\alpha}{2}}\right) + \mathrm{P}\left(\frac{\rho_{xg}\sqrt{n}(\hat{\beta}_{yx} - b + b)}{\sqrt{V}} > z_{\frac{\alpha}{2}}\right)$$

$$= \mathrm{P}\left(\frac{\rho_{xg}\sqrt{n}(\hat{\beta}_{yx} - b)}{\sqrt{V}} < -z_{\frac{\alpha}{2}} - \frac{b\rho_{xg}\sqrt{n}}{\sqrt{V}}\right) + \mathrm{P}\left(\frac{\rho_{xg}\sqrt{n}(\hat{\beta}_{yx} - b)}{\sqrt{V}} > z_{\frac{\alpha}{2}} - \frac{b\rho_{xg}\sqrt{n}}{\sqrt{V}}\right)$$

In the situation where β_{yx} is b, $\frac{\rho_{xg}\sqrt{n}(\hat{\beta}_{yx} - b)}{\sqrt{V}}$ has a standard normal distribution,

and so the last formula can be re-written with the cumulative distribution

function of the standard Normal distribution as

$$\Phi\left(-z_{\frac{\alpha}{2}} - \frac{b\rho_{xg}\sqrt{n}}{\sqrt{V}}\right) + \left[1 - \Phi\left(\frac{\rho_{xg}\sqrt{n}(\hat{\beta}_{yx} - b)}{\sqrt{V}} > z_{\frac{\alpha}{2}} - \frac{b\rho_{xg}\sqrt{n}}{\sqrt{V}}\right)\right]$$

Re-arranging provides the equation as given in the paper.

Appendix 2

We simulated the variables G, U, X and Y, representing the values of the SNP rs1205, confounders, log(CRP) and fibrinogen respectively, according to the distributions given here, to ensure that G would be a valid instrumental variable and that b, ρ_{xg}^2, V(X) and V(Y|X) would be equal to 0.234, 0.01, 1.11 and 1.11

respectively. This was done by letting G and U have marginal distributions as given in equations (8) and (9), and then simulating X with equation (10) in order for it to have the appropriate conditional distribution using standard probability theory, as given for example in Section 5.2 of DasGupta[24]. The variable *p* varied the level of confounding between X and Y due to U, and we simulated at p = 0.05, 0.5 and 0.95, corresponding to low, medium and high levels of confounding respectively. Equation (11) ensures that G only affects Y through X, that U and X have no interaction, and that Var(Y|X) equals 1.11.

$$\mathrm{G} \sim 1, 2, \text{or } 3 \text{ with probabilities } \frac{1}{9}, \frac{4}{9}, \frac{4}{9} \text{ respectively} \tag{8}$$

$$\mathrm{U} \sim \mathrm{N}(0, 1.11 \times p \times (1 - 0.01)) \tag{9}$$

$$\mathrm{X} \sim \mathrm{N}\left(\mathrm{U} + 0.1(\mathrm{G} - 2)\sqrt{1.11 \times \frac{9}{4}}, 1.11 \times (1 - p) \times (1 - 0.01)\right) \tag{10}$$

$$\mathrm{Y} = 0.234\mathrm{X} + \mathrm{U}\sqrt{\frac{1}{1 - 0.01}} \tag{11}$$

Table 1: Sample sizes required for a Mendelian randomization study for a range of parameter values and effect sizes when the p-value required is 0.05 and the power desired is 0.8, where b is the effect size that needs to be detected, ρ^2_{xg} is the square of the correlation between the genetic instrument and the exposure, and V is the ratio of the variance of the residuals after removing the effect of the exposure to the variance of the exposure itself.

V	**0.2**			**0.5**			**1**			**2**		
ρ^2_{xg}	**0.1**	**0.03**	**0.01**	**0.1**	**0.03**	**0.01**	**0.1**	**0.03**	**0.01**	**0.1**	**0.03**	**0.01**
b												
0.05	6,280	20,931	62,791	15,698	52,326	156,976	31,396	104,651	313,951	62,791	209,301	627,902
0.1	1,570	5,233	15,698	3,925	13,082	39,244	7,849	26,163	78,488	15,698	52,326	156,976
0.25	252	838	2,512	628	2,094	6,280	1,256	4,187	12,559	2,512	8,373	25,117
0.5	63	210	628	157	524	1,570	314	1,047	3,140	628	2,094	6,280

FIGURE LEGENDS

Figure 1. Causal graph showing the nature of the causal relationships between the genotype (G), exposure (X), unobserved confounders (U) and outcome (Y).

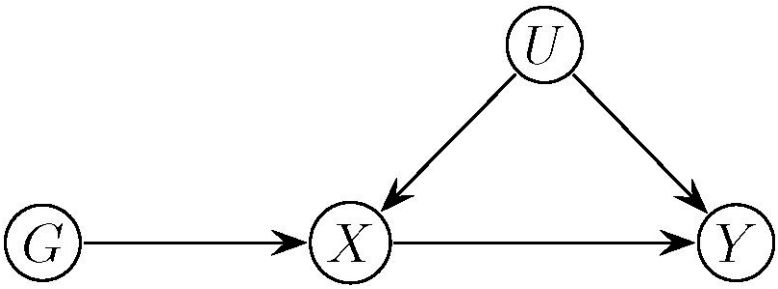

Figure 2. Contour plots showing minimum sample sizes required to obtain p-values of less than 0.05 with power 0.8 for different effect sizes for a range of correlations between the exposure and the genetic instrument at four levels of V $= \frac{Var(Y|X)}{Var(X)}$.